Cost Rate Heuristics for Semi-Markov Decision Processes

Glazebrook, K.D.; Bailey, Michael P.; Whitaker, Lyn R.

Applied Probability Trust

Glazebrook, K. D., Michael P. Bailey, and Lyn R. Whitaker. "Cost rate heuristics for semi-Markov decision processes." Journal of applied probability 29.3 (1992): 633-644.
http://hdl.handle.net/10945/63261

J. Appl. Prob. **29**, 633–644 (1992)
Printed in Israel

COST RATE HEURISTICS FOR SEMI-MARKOV DECISION PROCESSES

K. D. GLAZEBROOK,* *University of Newcastle upon Tyne*
MICHAEL P. BAILEY,** *Naval Postgraduate School, Monterey*
LYN R. WHITAKER,** *Naval Postgraduate School, Monterey*

Abstract

In response to the computational complexity of the dynamic programming/backwards induction approach to the development of optimal policies for semi-Markov decision processes, we propose a class of heuristics resulting from an inductive process which proceeds forwards in time. These heuristics always choose actions in such a way as to minimize some measure of the current cost rate. We describe a procedure for calculating such cost rate heuristics. The quality of the performance of such policies is related to the speed of evolution (in a cost sense) of the process. A simple model of preventive maintenance is described in detail. Cost rate heuristics for this problem are calculated and assessed computationally.

DYNAMIC PROGRAMMING; REPLACEMENT POLICY

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 90C39

1. Introduction

Much research in discounted Markov and semi-Markov decision processes has centred around efficient implementations of value iteration (see Howard (1960)). Many authors (see Porteus (1980) for an overview) have studied refinements to the basic scheme. This large body of work is motivated, *inter alia*, by the inherent computational complexity of the dynamic programming/backwards induction approach. See Ross (1970) for an accessible account of iterative schemes for the solution of the semi-Markov decision processes of primary interest here.

Gittins (1989) describes an interestingly novel approach to the construction of policies for discounted semi-Markov decision processes. At time 0, a policy (π_1, say) and a stopping time on the process under π_1 (τ_1, say) are chosen to minimize a natural measure of cost rate incurred from the initial state at 0 up to the stopping time. The forwards

Received 4 March 1991; revision received 20 June 1991.

* Postal address: Department of Mathematics and Statistics, The University, Newcastle upon Tyne NE1 7RU, UK.

Research supported by the National Research Council by means of a Senior Research Associateship at the Department of Operations Research, Naval Postgraduate School, Monterey, California.

** Postal address: Department of Operations Research, Naval Postgraduate School, Monterey, CA 93943, USA.

Dr Bailey was supported by the Naval Weapons Support Centre, Crane, IN, and Dr Whitaker by the Naval Postgraduate School Research Foundation.

induction policy constructed by this procedure then implements π_1 up to time τ_1. The state of the process at τ_1 ($X(\tau_1)$, say) is observed and a new policy/stopping time pair (π_2, τ_2, say) is chosen to minimize the cost rate from $X(\tau_1)$. Policy π_2 is then implemented during $[\tau_1, \tau_1 + \tau_2)$, and so on. Some strengths of this approach include the following:

(i) Forwards induction policies are optimal for a large class of models, especially in stochastic resource allocation. See Gittins (1989).

(ii) The on-line computation of such policies can often be performed in a way which offers considerable computational savings over conventional dynamic programming. See Katehakis and Veinott (1987) for a discussion.

(iii) The approach sometimes results in policies of simple structure (e.g. index-based). More generally it offers the prospect of relationships between model structure and policy structure which are theoretically accessible and (relatively) easily understood. See this illustrated in Glazebrook (1991).

We propose a general approach to the development of heuristics for discounted semi-Markov decision processes which uses cost rates in a simpler fashion than in forwards induction, but which retains some of that procedure's strengths — especially those mentioned under (ii) and (iii) above. The approach is quasi-myopic and offers particular advantages in situations where to assume a fixed stationary model over an infinite horizon would be hazardous. In these heuristics, a simple choice for the stopping times τ_n, $n \geqq 1$, is made *a priori* and cost rate minimizations are over policies only. This class of cost rate heuristics is introduced in Section 2, together with a procedure for their computation. Performance bounds for these heuristics are developed in Section 3. These ideas are illustrated in Section 4 by means of a computational assessment of cost rate heuristics for a simple machine replacement problem. A cost rate myopic policy is found to perform well much of the time.

2. Cost rate heuristics for semi-Markov decision processes

Our model is a discounted semi-Markov decision process (see for instance Ross (1970)) with the following special features:

(i) *States*. $X(t)$ denotes the state of the process at time $t \in \mathscr{R}_{\geqq 0}$. The state space Ω is a Borel subset of some complete separable metric space, together with a σ-algebra $\mathscr{F}$ of subsets of Ω which includes every single element subset.

(ii) *Actions*. At every decision epoch t one of the actions $a_1, a_2, \cdots, a_N$ is taken where $N < \infty$. A *stationary policy* is a map $\pi : \Omega \to \{a_1, a_2, \cdots, a_N\}$, the interpretation being that policy π takes action $\pi\{X(t)\}$ at time t. In general, a *policy* is any rule for choosing actions (satisfying the obvious measurability requirements) which is a function of the history of the process up to the current decision epoch. Such policies may be randomised.

(iii) *Costs*. If action a_j is taken at decision epoch t an expected cost $\alpha^t c\{X(t), a_j\}$ is incurred. Here $\alpha \in [0, 1)$ is the *discount rate* and for each a_j, $c(\cdot, a_j) : \Omega \to \mathscr{R}_{\geqq 0}$ is a bounded, $\mathscr{F}$-measurable function.

(iv) *Process evolution*. If action a_j is taken at decision epoch t then
(a) a state transition is observed, and
(b) a random amount of time elapses before the next decision epoch.

$\boldsymbol{P}(G \mid x, a_j)$ is the probability that the state of the process at the next epoch lies in set $G \in \mathcal{F}$ conditional upon the event $X(t) = x$. $\boldsymbol{F}(H \mid x, y, a_j)$ is the probability that the time to the next decision epoch lies in *Borel set* H given that a transition from $x\,(= X(t))$ to y occurs. $\boldsymbol{P}(G \mid \cdot, a_j) : \Omega \to [0, 1]$ is $\mathcal{F}$-measurable and $\boldsymbol{F}(H \mid \cdot, \cdot, a_j) : \Omega \times \Omega \to [0, 1]$ is $\mathcal{F} \times \mathcal{F}$-measurable. We shall denote by $\boldsymbol{P}^r$, $\boldsymbol{F}^r$ the equivalent r-step measures — e.g. $\boldsymbol{P}^r(G \mid x, \pi)$ is the probability that the state of the process at the rth decision epoch after t lies in set G, given that $X(t) = x$ and that policy π (assumed not to depend upon the history of the process before t) is adopted. The first decision epoch is always assumed to be 0.

The following condition is standard in the study of semi-Markov decision processes (see for instance Ross (1970)). It guarantees (with probability 1) that we do not have an infinite number of decision epochs in finite time.

Condition 1. There exist positive ε, δ such that

$$\int_{\Omega} \boldsymbol{F}\{(\delta, \infty) \mid x, y, a_j\} \boldsymbol{P}(dy \mid x, a_j) > \varepsilon, \qquad 1 \leqq j \leqq N, \quad x \in \Omega.$$

(v) *Optimal policies*. Denote by $C_r(\pi, x)$ the total expected cost incurred from the imposition of policy π from time 0 for r decision epochs when $X(0) = x$. If π is stationary $C_r(\pi, \cdot)$ may be recovered from the recursion:

$$C_0(\pi, x) = 0;$$

$$C_r(\pi, x) = c\{x, \pi(x)\} + \int_{\Omega} \int_{t=0}^{\infty} \alpha^t C_{r-1}(\pi, y) \boldsymbol{F}\{dt \mid x, y, \pi(x)\} \boldsymbol{P}\{dy \mid x, \pi(x)\}, \quad r \geqq 1.$$

We define

$$C(\pi, x) \equiv \lim_{r \to \infty} C_r(\pi, x) \tag{1}$$

as the total expected cost incurred by policy π when $X(0) = x$. The above assumptions (in particular the boundedness of costs and Condition 1) guarantee not only that the limit in (1) exists, but that the convergence is uniform over all policies π, for all $x \in \Omega$.

A policy π^* is *optimal* if

$$C(\pi^*, x) = \inf_{\pi} C(\pi, x) \equiv C(x), \qquad x \in \Omega.$$

The general theory (see Blackwell (1965)) asserts the existence of an optimal policy π^* which is stationary and such that $C(\cdot)$ uniquely satisfies the recursion

$$C(x) = \min_{1 \leqq j \leqq N} \left\{ c(x, a_j) + \int_{\Omega} \int_{t=0}^{\infty} \alpha^t C(y) \boldsymbol{F}(dt \mid x, y, a_j) \boldsymbol{P}(dy \mid x, a_j) \right\}. \tag{2}$$

Procedures for determining $C(\cdot)$ and π^* include *value iteration* and *policy iteration*, as described by Ross (1970).

Now, write $\tau_r(\pi, x)$ for the random time of the rth decision epoch after 0 when policy π is adopted and $X(0) = x$. We write $M_r(\pi, x) \equiv \boldsymbol{E}\{\alpha^{\tau_r(\pi,x)}\}$. If π is stationary $M_r(\pi, \cdot)$ may be recovered from the recursion

$$M_0(\pi, x) = 1;$$
$$M_r(\pi, x) = \int_\Omega \int_{t=0}^{\infty} \alpha^t M_{r-1}(\pi, y) \boldsymbol{F}\{dt \mid x, y, \pi(x)\} \boldsymbol{P}\{dy \mid x, \pi(x)\}, \quad r \geqq 1.$$

Note that Condition 1 guarantees that for all π, x

$$1 > (1 - \varepsilon + \varepsilon\alpha^\delta)^r \geqq M_r(\pi, x), \qquad r \geqq 1. \tag{3}$$

The notion expressed in Definition 1 is central to the ideas explored in the paper.

Definition 1. The *r-stage cost rate function* for policy π, $\Gamma_r(\pi, \cdot) : \Omega \to \mathscr{R}_{\geqq 0}$ is given by

$$\Gamma_r(\pi, x) \equiv C_r(\pi, x)\{1 - M_r(\pi, x)\}^{-1}. \tag{4}$$

The rationale for calling $\Gamma_r(\pi, x)$ a cost rate emerges from the identity

$$\Gamma_r(\pi, x) \equiv C_r(\pi, x)\left[\boldsymbol{E}\left\{\int_0^{\tau_r(\pi,x)} \alpha^t dt\right\}\right]^{-1} (-\ln \alpha)^{-1}, \tag{5}$$

in which the notion of averaging is an (appropriately) discounted one.

Definition 2. Policy $\hat{\pi}$ is (r, x)-*optimal* (r-stage cost rate optimal for state x) if

$$\Gamma_r(\hat{\pi}, x) = \inf_\pi \Gamma_r(\pi, x). \tag{6}$$

In order to explore the properties of r-stage cost rates and associated optimal policies we introduce the mapping $T_r(x, \cdot) : \mathscr{R}_{\geqq 0} \to \mathscr{R}_{\geqq 0}$ defined by

$$T_r(x, u) = \inf_\pi \{C_r(\pi, x) + uM_r(\pi, x)\} \tag{7}$$

and its n-fold version $T_r^n(x, \cdot) : \mathscr{R}_{\geqq 0} \to \mathscr{R}_{\geqq 0}$, where

$$T_r^n(x, u) = T_r\{x, T_r^{n-1}(x, u)\}, \qquad n \geqq 1.$$

Equation (7) defines a finite horizon dynamic program. We may assert the existence of a policy $\pi : \Omega \times \{1, 2, \cdots, r\} \to \{a_1, a_2, \cdots, a_N\}$ attaining the infimum in (7). Here $\pi(x, s)$ is the action taken by policy π when in state $x \in \Omega$ at the sth decision epoch. Call such a policy *r-stage stationary*.

Theorem 1. For each $x \in \Omega$, $r \geqq 1$,

(a) $T_r(x, \cdot)$ is monotonic, non-decreasing;

(b) $T_r(x, \cdot)$ is a contraction mapping;

(c) $\Gamma = \inf_\pi \Gamma_r(\pi, x)$ is the unique member of $\mathscr{R}_{\geqq 0}$ for which

$$T_r(x, \Gamma) = \Gamma;$$

(d) there exists an (r, x)-optimal policy which is r-stage stationary;

(e) for each $u \in \mathscr{R}_{\geqq 0}$

$$\lim_{n\to\infty} T_r^n(x, u) = \Gamma = \inf_{\pi} \Gamma_r(\pi, x),$$

this convergence being geometrical and uniform over x.

Proof.

(a) It is trivial from (7) that $u \geqq v \Rightarrow T_r(x, u) \geqq T_r(x, v)$.

(b) Suppose that $u \geqq v$. Write $\pi(u)$ for an r-stage stationary policy attaining the infimum in (7). It is plain that

$$0 \leqq T_r(x, u) - T_r(x, v) \leqq M_r\{\pi(v), x\}(u - v) \leqq (1 - \varepsilon + \varepsilon\alpha^{\delta})^r(u - v),$$

from (3). This establishes (b).

(c) The contraction mapping fixed point theorem guarantees the existence of a unique fixed point for $T_r(x, \cdot)$. Call the fixed point γ. Write

$$\gamma = T_r(x, \gamma) = \inf_{\pi} \{C_r(\pi, x) + \gamma M_r(\pi, x)\} = C_r\{\pi(\gamma), x\} + \gamma M_r\{\pi(\gamma), x\} \tag{8}$$

where we write $\pi(\gamma)$ for a policy attaining the infimum in (8). It now follows that

$$\gamma = C_r\{\pi(\gamma), x\}[1 - M_r\{\pi(\gamma), x\}]^{-1} = \Gamma_r\{\pi(\gamma), x\} \geqq \Gamma.$$

Suppose that $\gamma > \Gamma$, and obtain a contradiction. We now have a policy $\tilde{\pi}$, say, such that $\gamma > C_r(\tilde{\pi}, x)\{1 - M_r(\tilde{\pi}, x)\}^{-1}$ from which it follows that

$$\begin{aligned} \gamma &> C_r(\tilde{\pi}, x) + \gamma M_r(\tilde{\pi}, x) \\ &\geqq \inf_{\pi} \{C_r(\pi, x) + \gamma M_r(\pi, x)\} \Rightarrow \gamma > T_r(x, \gamma), \end{aligned}$$

from which we conclude that γ is not a fixed point of $T_r(x, \cdot)$, a contradiction. Hence $\gamma = \Gamma$, and we have established (c).

(d) It is now plain that any policy $\pi(\Gamma)$ attaining the infimum in (7) with $u = \Gamma$ is (r, x)-optimal. We have already noted that there is one such which is r-stage stationary. We have proved the result.

(e) This is a standard consequence of (b) and (c).

The above result plainly yields a value iteration approach to the computation of minimal cost rates and hence of (r, x)-optimal policies. We now describe the class of cost rate heuristics for semi-Markov decision processes of primary interest to us. In Definition 3, $\boldsymbol{r} \equiv [\{r_n(\cdot) : \Omega \to \mathbb{Z}^+\}, n \in \mathbb{Z}^+]$ is a sequence of $\mathscr{F}$-measurable functions taking values in the positive integers.

Definition 3. A *cost rate heuristic* determined by $\boldsymbol{r}$ is denoted $\hat{\pi}(\boldsymbol{r})$ and is a policy which operates as follows:

(a) If $X(0) = x$, $\hat{\pi}(\boldsymbol{r})$ takes the first $r_1(x)$ decisions according to an $\{r_1(x), x\}$-optimal policy;

(b) Suppose that the state of the process following the first $\Sigma_{m=1}^{n} r_m(X_{m-1})$ decisions and transitions (i.e. following the first n stages) under policy $\hat{\pi}(\boldsymbol{r})$ is X_n, $n \geqq 1$, where

$X_0 \equiv X(0)$. Policy $\hat{\pi}(\boldsymbol{r})$ takes the next $r_{n+1}(X_n)$ decisions according to an $\{r_{n+1}(X_n), X_n\}$-optimal policy, $n \geqq 1$.

Comments.

1. Hence policy $\hat{\pi}(\boldsymbol{r})$ implements an $\{r_1(x), x\}$-optimal policy from time 0 when $X(0) = x$ as a procedure for determining the first $r_1(x)$ decisions. The state is then updated to X_1. The number of decisions to be taken in the second stage is $r_2(X_1)$ and is allowed to depend upon X_1. An $\{r_2(X_1), X_1\}$-optimal policy is computed and implemented from state X_1, and so on.

2. Apart from any possibility there might be of obtaining (r, x)-optimal policies of special structure, a major opportunity for cost rate heuristics to reduce computational requirements (as compared with the application of standard dynamic programming) arises from the fact that value iteration for (r, x)-optimal policies based on Theorem 1 only needs to look at states which are accessible in r steps from state x. In Bayesian sequential problems, an example of which is described in Section 4, considerable savings are often possible. Another instance is where state variable x is enhanced to include (for example) the number of decisions taken to date as a means of accommodating non-stationarity.

3. If each function $r_n(\cdot)$ is a constant (i.e., the number of decisions in each stage is fixed at the outset), $\hat{\pi}(\boldsymbol{r})$ is called a *fixed sequence cost rate heuristic*. We shall often be interested in fixed sequence policies for which $r_n(\cdot) \equiv 1$, $n \geqq 2$. In relation to such a choice note that $(1, x)$-optimal policies are often trivial to compute. Cost rate heuristics for which $r_n(\cdot) \equiv 1$, $n \geqq 1$, will be called *cost rate myopic*.

We now explore further the rationale for considering such heuristics.

3. General performance bounds for cost rate heuristics

Write $\Delta(x, y) = C(y) - C(x) \equiv C(\pi^*, y) - C(\pi^*, x)$ for the change in minimal costs which occurs upon a transition from x to y. As before, write $\tau_r(\pi, x)$ for the random time of the rth decision epoch after 0 when policy π is adopted and $X(0) = x$. The subscript in the notation $\boldsymbol{E}_\pi$ indicates that an expectation is to be taken over realisations of the system conditional upon implementation of the policy π.

Definition 4. The *r-decision speed function* for policy π, $\Delta_r(\pi, \cdot) : \Omega \to \mathscr{R}$ is given by

$$\begin{aligned}\Delta_r(\pi, x) &\equiv \boldsymbol{E}_\pi\{\alpha^{\tau_r(\pi,x)}\Delta[x, X\{\tau_r(\pi, x)\}]\}\{1 - M_r(\pi, x)\}^{-1} \\ &= \left[\left\{\int_\Omega \int_{t=0}^{\infty} \alpha^t C(y) F^r(dt \mid x, y, \pi) P^r(dy \mid x, \pi)\right\} - M_r(\pi, x) C(x)\right] \\ &\quad \times \{1 - M_r(\pi, x)\}^{-1}. \end{aligned} \tag{9}$$

See (9): $\Delta_r(\pi, x)$ represents a (discounted) rate at which future prospects (as measured by $C(\cdot)$) change during an r-decision implementation of policy π. It will emerge that we can go some way towards analysing policies in terms of a combination of cost rate and speed functions. The following result is an example.

Lemma 2. For each $x \in \Omega$, $r \geqq 1$, $C(x) = \Gamma_r(\pi^*, x) + \Delta_r(\pi^*, x)$.

Proof. Recall that $C(x) = C(\pi^*, x) = \inf_\pi C(\pi, x)$. By standard results, $C(\cdot)$ satisfies the recursion

$$\begin{aligned} C(x) &= C(\pi^*, x) \\ &= C_r(\pi^*, x) + \int_\Omega \int_{t=0}^{\infty} \alpha^t C(y) \boldsymbol{F}^r(dt \mid x, y, \pi^*) \boldsymbol{P}^r(dy \mid x, \pi^*) \\ &= \Gamma_r(\pi^*, x)\{1 - M_r(\pi^*, x)\} + \Delta_r(\pi^*, x)\{1 - M_r(\pi^*, x)\} + M_r(\pi^*, x) C(x) \end{aligned}$$

from (4) and (9). Invoking (3), the result follows trivially.

Lemma 3. For each π and $x \in \Omega$

$$\lim_{r \to \infty} \Delta_r(\pi, x) = 0, \tag{10}$$

the convergence in (10) being uniform over all policies π and states x.

Proof. From (3) and (9)

$$|\Delta_r(\pi, x)| \leqq \left\{ \sup_{x \in \Omega} C(x) \right\} (1 - \varepsilon + \varepsilon \alpha^\delta)^r \{1 - (1 - \varepsilon + \varepsilon \alpha^\delta)^r\}^{-1}. \tag{11}$$

The result follows trivially.

Lemmas 2 and 3 create the expectation that (crudely speaking) should a decision process have uniformly small r-decision speed functions then an analysis in terms of r-decision cost rates could be successful. Lemma 3 tells us that we can always force the speed functions to be small by choosing r large enough. However, we note that the larger r is, the more computationally demanding is the development of (r, x)-optimal policies. We make these ideas more explicit as follows. Suppose that $\hat{\pi}$ is an (r, x)-optimal policy (see Theorem 1(d)). Write

$$C^r(x) = C_r(\hat{\pi}, x) + \int_\Omega \int_{t=0}^{\infty} \alpha^t C(y) \boldsymbol{F}^r(dt \mid x, y, \hat{\pi}) \boldsymbol{P}^r(dy \mid x, \hat{\pi}) \tag{12}$$

for the total expected cost from implementing $\hat{\pi}$ for r decisions, thereafter followed by an optimal policy. Theorem 4 bounds how much is lost by pursuing $\hat{\pi}$ instead of an optimal policy for these first r decisions.

Theorem 4. For each $x \in \Omega$, $r \geqq 1$,

$$C^r(x) - C(x) \leqq \{\Delta_r(\hat{\pi}, x) - \Delta_r(\pi^*, x)\}\{1 - M_r(\hat{\pi}, x)\} \to 0, \quad \text{as } r \to \infty, \tag{13}$$

uniformly over all states x.

Proof. From (9) and (12),

$$\begin{aligned} C^r(x) = C_r(\hat{\pi}, x) + \int_\Omega \int_{t=0}^{\infty} \alpha^t [\Delta(x, y) + \{C(x) - C^r(x)\} + C^r(x)] \\ \times \boldsymbol{F}^r(dt \mid x, y, \hat{\pi}) \boldsymbol{P}^r(dy \mid x, \hat{\pi}) \end{aligned}$$

$$= C_r(\hat{\pi}, x) + \Delta_r(\hat{\pi}, x)\{1 - M_r(\hat{\pi}, x)\}$$
$$+ \{C(x) - C^r(x)\}M_r(\hat{\pi}, x) + C^r(x)M_r(\hat{\pi}, x).$$

Hence we deduce that

$$C^r(x) = \Gamma_r(\hat{\pi}, x) + \Delta_r(\hat{\pi}, x) + \{C(x) - C^r(x)\}M_r(\hat{\pi}, x)\{1 - M_r(\hat{\pi}, x)\}^{-1}.$$

Now, from Lemma 2

$$C^r(x) - C(x) = \{\Gamma_r(\hat{\pi}, x) - \Gamma_r(\pi^*, x)\} + \{\Delta_r(\hat{\pi}, x) - \Delta_r(\pi^*, x)\}$$
$$+ \{C(x) - C^r(x)\}M_r(\hat{\pi}, x)\{1 - M_r(\hat{\pi}, x)\}^{-1}$$
$$\leqq \Delta_r(\hat{\pi}, x) - \Delta_r(\pi^*, x) + \{C(x) - C^r(x)\}M_r(\hat{\pi}, x)\{1 - M_r(\hat{\pi}, x)\}^{-1},$$

since $\hat{\pi}$ is (r, x)-optimal and so $\Gamma_r(\hat{\pi}, x) \leqq \Gamma_r(\pi^*, x)$. Inequality (13) now follows trivially. The convergence result is a simple consequence of Lemma 3.

It is in fact possible to deduce from Theorem 4 a bound on the suboptimality of $\hat{\pi}(\boldsymbol{r})$ expressed in terms of speed functions by a suitable aggregation of quantities like that on the right-hand side of (13).

We now proceed to a computational study of the performance of cost rate heuristics for a simple replacement problem.

4. A simple model of preventative maintenance

A system is subject to random deterioration and failure. A new system is installed at time 0 and (in the absence of intervention) its time to failure has distribution F_θ where $\theta \in \Theta$ is unknown. Replacing a failed system is expensive. At time t the cost is $\alpha^t c_1$ where as usual $\alpha \in [0, 1)$ is a discount rate. Alternatively, a (less expensive) planned replacement can be made in advance of system failure — here the cost at t is $\alpha^t c_2$.

Hence at time 0, one of N possible (planned) replacement times $0 < a_1 < a_2 < \cdots < a_N$ must be chosen. Note that we might have $a_N = \infty$, i.e. the choice of such an a_N implies that the system is left to fail with no planned replacement in anticipation of failure. We have $X_1, X_2, \cdots$ a sequence of i.i.d. system failure times with $X_i \sim F_\theta$. If action a_i is taken at 0, a planned replacement occurs at a_i if $X_1 > a_i$ and otherwise the system is replaced at failure. At time $\min(X_1, a_i)$ one of the N replacement times $\{a_j, 1 \leqq j \leqq N\}$ is chosen for the new system. We proceed in this fashion. Choosing replacement times which are too small incurs unnecessary costs from a surfeit of planned replacements. Replacement times which are too large carry the risk of large numbers of expensive replacements upon failure of the system. We suppose that θ has a prior distribution G and look for a Bayes sequential decision rule for this problem.

This problem (in common with, say, bandit problems) presents in a simple way the tension between taking decisions whose prime purpose is to gain information (and hence improve the quality of future decisions) and taking decisions which exploit information already available.

More elaborate versions of this problem are discussed for models with known stochastic structure (i.e. known θ) by Aven (1983) and Chen and Savits (1988), both of whom use one-stage cost rates in their analysis. Attempts at learning about such a system have usually been structured according to partially observable Markov decision processes — see Albright (1978) and White (1979). In models with the average cost per unit time criterion, Bather (1977), Frees and Ruppert (1985) and Aras and Whitaker (1990) have taken non-Bayesian and non-parametric approaches to learning about the underlying system.

In our Bayesian model a cost rate myopic policy chooses a_i to minimize

$$\left[\int_{\Theta}\left(\int_0^{a_i}\alpha^t c_1 F_\theta(dt)+\alpha^{a_i}c_2 F_\theta\{[a_i,\infty)\}\right)\tilde{G}(d\theta)\right] \times\left(1-\int_{\Theta}\left(\int_0^{a_i}\alpha^t F_0(dt)+\alpha^{a_i}F_\theta\{[a_i,\infty)\}\right)\tilde{G}(d\theta)\right)^{-1} \tag{14}$$

where $\tilde{G}$ is the current posterior for θ. Hence cost rate myopic policies are adaptive, depending as they do upon the current posterior. Executing the minimization in (14) is usually computationally trivial, rendering this class of policies attractive as heuristics.

In our discussion of cost rate heuristics for this problem we shall study only those $\hat{\pi}(\boldsymbol{r})$ with $r_n = 1$, $n \geqq 2$. In fact it is a simple consequence of Lemma 3 and Theorem 4 that for any $\gamma > 0$ and $x \in \Omega$ we can always ensure that $C(\hat{\pi}(\boldsymbol{r}), x) - C(x) \leqq \gamma$ by choosing r_1 sufficiently large. Again from Theorem 4 we see that the important question of the smallest r_1 yielding acceptable performance in $\hat{\pi}(\boldsymbol{r})$ is related to values of the appropriate speed functions. It would seem intuitive that in the current Bayesian context, speed functions for policies should be related to the spread (loosely defined) of the current posterior. For example, if the prior G for real-valued θ has small variance (i.e. we are close to the known θ case) we would expect all speed functions to be small and the cost rate myopic policy with $r_1 = 1$ to perform well. Less precise prior knowledge may necessitate a choice of $r_1 > 1$ to allow for some initial learning. We now present some computational results which illustrate these phenomena.

We consider a replacement problem with $c_1 = 10$, $c_2 = 1$ and $\alpha = 0.99$. Failure times are assumed to be independent Weibull $(n, 0.4)$ random variables, i.e. having density

$$f(x; n, \lambda) = \lambda n x^{n-1}\exp(-\lambda x^n), \qquad x > 0$$

with $\lambda = 0.4$. G is a two-point prior with

$$G(n_1) = p = 1 - G(n_2) \tag{15}$$

where $n_1 = 1$ and $n_2 = 8$. At each decision epoch we are faced with a choice between $N = 50$ planned replacement times given by

$$a_j = 1.0 + (j-1)0.04, \qquad 1 \leqq j \leqq 50.$$

For simplicity of notation, denote by $C(p)$ the Bayes cost incurred when adopting an optimal policy with prior distribution (15) and $C_m(p)$ the equivalent cost from adopting $\hat{\pi}(\boldsymbol{r})$ with $r_1 = m$; $r_n = 1$, $n \geqq 2$. The (m, p)-optimal policy which constitutes the first

stage of $\hat{\pi}(\boldsymbol{r})$ is calculated according to the computational procedure derived from Theorem 1. It may be of interest to note that in this procedure the number of calculations per iteration grows linearly in m. The computation of $(1, p)$-optimal policies is trivial. The costs $C(p)$, $C_m(p)$ are computed by value iteration or some simple variant of it.

In Table 1 find values of the relative percentage differences $100\{C_m(p) - C(p)\}\{C(p)\}^{-1}$, for $m = 1, 2, 3$, and $p = 0(0.1)1$. Values of $C(p)$ have been given in order that the absolute differences may be recovered.

TABLE 1
Relative percentage differences between the cost from heuristic $\hat{\pi}(\boldsymbol{r})$ and an optimal policy

p	$C(p)$	$100\{C_1(p) - C(p)\}\{C(p)\}^{-1}$	$100\{C_2(p) - C(p)\}\{C(p)\}^{-1}$	$100\{C_3(p) - C(p)\}\{C(p)\}^{-1}$
0.0	75.852	0.000	0.000	0.000
0.1	106.837	1.022	0.646	0.377
0.2	138.205	1.396	0.881	0.515
0.3	169.722	1.551	0.978	0.572
0.4	201.308	1.593	1.004	0.589
0.5	233.033	1.581	0.995	0.586
0.6	264.949	1.506	0.949	0.557
0.7	297.128	1.360	0.861	0.499
0.8	329.797	1.136	0.721	0.414
0.9	363.534	0.735	0.475	0.260
1.0	400.393	0.000	0.000	0.000

Figure 1 presents these data graphically.

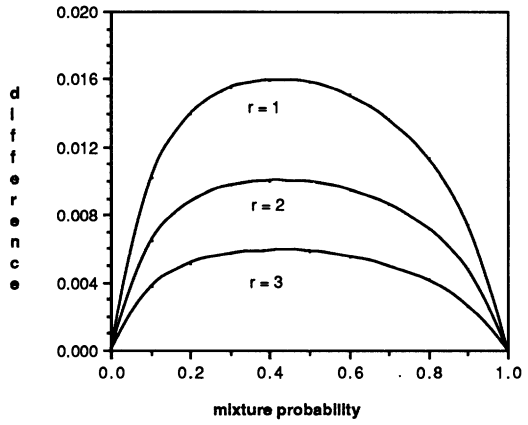

Figure 1. Relative percentage differences between the cost from heuristic $\hat{\pi}(\boldsymbol{r})$ and an optimal policy

If, for example, we wish to choose a heuristic $\hat{\pi}(\boldsymbol{r})$ whose Bayes' cost is within 1% of the optimum then, from Table 1, choosing $r_1 = 1$ would suffice for $p = 0, 0.9, 1.0$; choosing $r_1 = 2$ would suffice for $p = 0.1, 0.2, 0.3, 0.5, 0.6, 0.7$ and 0.8 but we would need $r_1 = 3$ to attain this level of performance when $p = 0.4$. This pattern of behaviour is what the

above discussion (relating speed functions to the variance of the prior) would lead us to expect.

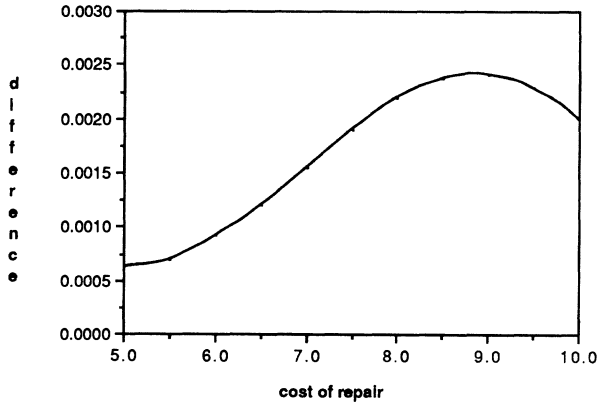

Figure 2. Absolute differences between the cost from the cost rate myopic policy and an optimal policy when $p = 0.1$

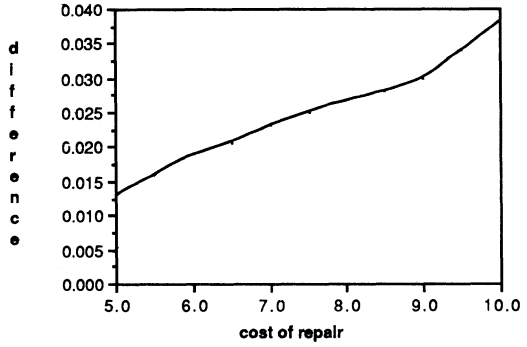

Figure 3. Absolute differences between the cost from the cost rate myopic policy and an optimal policy when $p = 0.5$

A striking feature of our numerical study of this replacement problem is the consistently strong performance of the cost rate myopic policy with $r_1 = 1$. In Figures 2, 3 and 4 find values of the absolute difference $C_1(p) - C(p)$ for the problem described above but with discount rate now taken to be $\alpha = 0.95$, a range of repair costs $c_1 = 5(0.5)10$ and c_2 fixed at 1. Figures 2 and 4 are for cases with small prior variance ($p = 0.1$ and 0.9 respectively) and Figure 3 for large prior variance. To give some idea of relative percentage differences in cost the ranges of $C(p)$ in Figures 2–4 are [15.75, 20.93], [26.57, 46.85] and [36.89, 72.56] respectively. Nowhere does the loss caused by adopting a cost rate myopic policy instead of an optimal policy exceed a fraction of 1%.

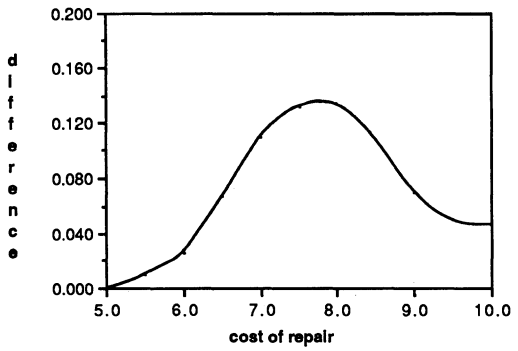

Figure 4. Absolute differences between the cost from the cost rate myopic policy and an optimal policy when $p = 0.9$

References

ALBRIGHT, S. C. (1978) Structural results for partially observable Markov decision processes. *Operat. Res.* **27**, 1041–1053.

ARAS, G. AND WHITAKER, L. R. (1990) *Sequential Nonparametric Estimation of an Age Replacement Policy*. Technical Report, Department of Operations Research, Naval Postgraduate School, Monterey, California.

AVEN, T. (1983) Optimal replacement under a minimal repair strategy — a general failure model. *Adv. Appl. Prob.* **15**, 198–211.

AVEN, T. AND BERGMAN, B. (1986) Optimal replacement times — a general setup. *J. Appl. Prob.* **23**, 432–442.

BATHER, J. A. (1977) On the sequential construction of an optimal age replacement policy. *Bull. Inst. Internat. Statist.* **47**, 253–266.

BLACKWELL, D. (1965) Discounted dynamic programming. *Ann. Math. Statist.* **36**, 226–235.

CHEN, C. S. AND SAVITS, T. H. (1988) A discounted cost relationship. *J. Multivariate Anal.* **27**, 105–115.

FREES, E. AND RUPPERT, D. (1985) Sequential nonparametric age replacement policies. *Ann. Statist.* **13**, 650–662.

GITTINS, J. C. (1989) *Multi-Armed Bandit Allocation Indices*. Wiley, Chichester.

GLAZEBROOK, K. D. (1991) Strategy evaluation for stochastic scheduling problems with order constraints. *Adv. Appl. Prob.* **23**, 86–104.

HOWARD, R. (1960) *Dynamic Programming and Markov Processes*. MIT Press, Cambridge, MA.

KATEHAKIS, M. N. AND VEINOTT, A. F. (1987) The multi-armed bandit problem: decomposition and computation. *Math. Operat. Res.* **12**, 262–268.

PORTEUS, E. (1980) Overview of iterative methods for discounted finite Markov and semi-Markov decision chains. In *Recent Developments in Markov Decision Processes*, ed. R. Hartley, L. C. Thomas and D. J. White, 1–20.

ROSS, S. M. (1970) *Applied Probability Models with Optimization Applications*. Holden-Day, San Francisco.

WHITE, C. C. (1979) Bounds on optimal cost for a replacement problem with partial observations. *Nav. Res. Logist. Quart.* **26**, 415–422.